Diffusion theory-based models of demographic history

Sergio Lukic

Rutgers University

1.0. Introduction

Newly emerging DNA sequence data sets offer the potential to reveal the demographic histories of populations, as well as the role of natural selection on individual genes and populations.

However, extracting this kind of information requires mathematical models that can capture the diversity and richness of population histories, and efficient tools that can fit models to massive amounts of data.

1.0. Introduction

In populations of the same species, statistical inference based on the distribution of allele frequencies is the preferred approach.

The evolution of allele frequencies depends on demography as well as on natural selection.

Recent population growth can be misunderstood as a signature of purifying selection.

Recent population admixture can cause non-trivial patterns of Linkage Disequilibrium, which can be misinterpreted as hitchhiking effects and/or epistasis.

1.0. Introduction

Complex demographic histories that involve several population splitting events, gene-flow, and population bottlenecks commonly underlie patterns of genetic variation in humans.

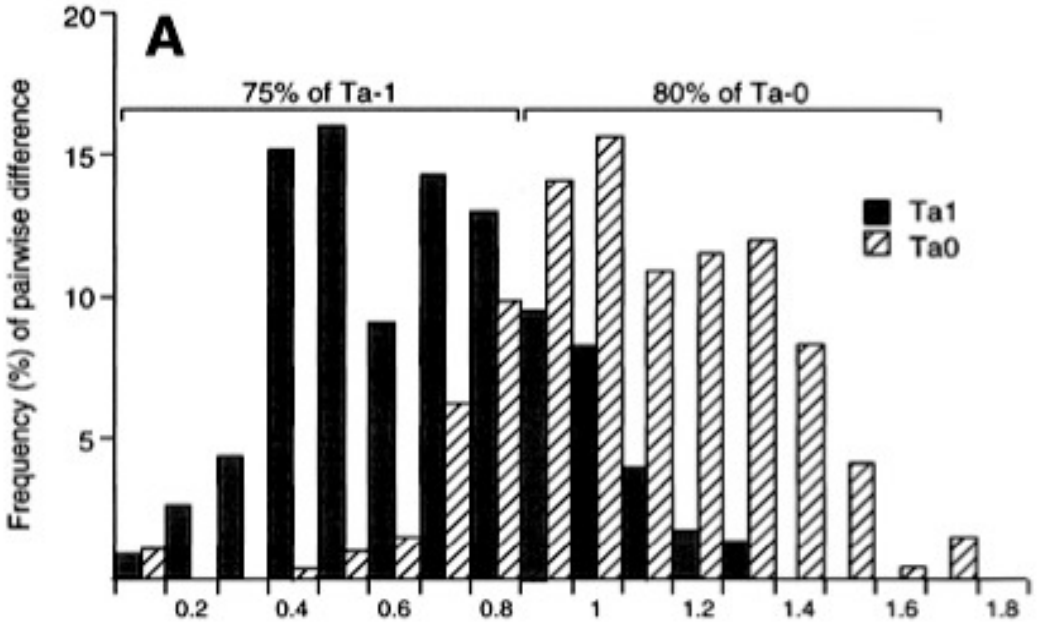
Most existing methods for studying demography are limited to working with relatively *small numbers of loci and sampled chromosomes*.

A growing amount of genome-wide sequence and polymorphism data motivates the development of new tools for the study of demography.

1.1. One motivation

Element	Chimpanzee	Human
Alu	2,340 (0.7Mb)	7,082 (2.1 Mb)
LINE-1	1,979 (5.5Mb)	1,814 (5.0 Mb)
SVA	757 (1.1 Mb)	970 (1.3 Mb)
ERV class 1	234 (.1Mb)	5 (8 kb)
ERV class 2	45 (55 kb)	77 (130 kb)

Chimpanzee Sequencing and Analysis Consortium. *Nature* (2005)

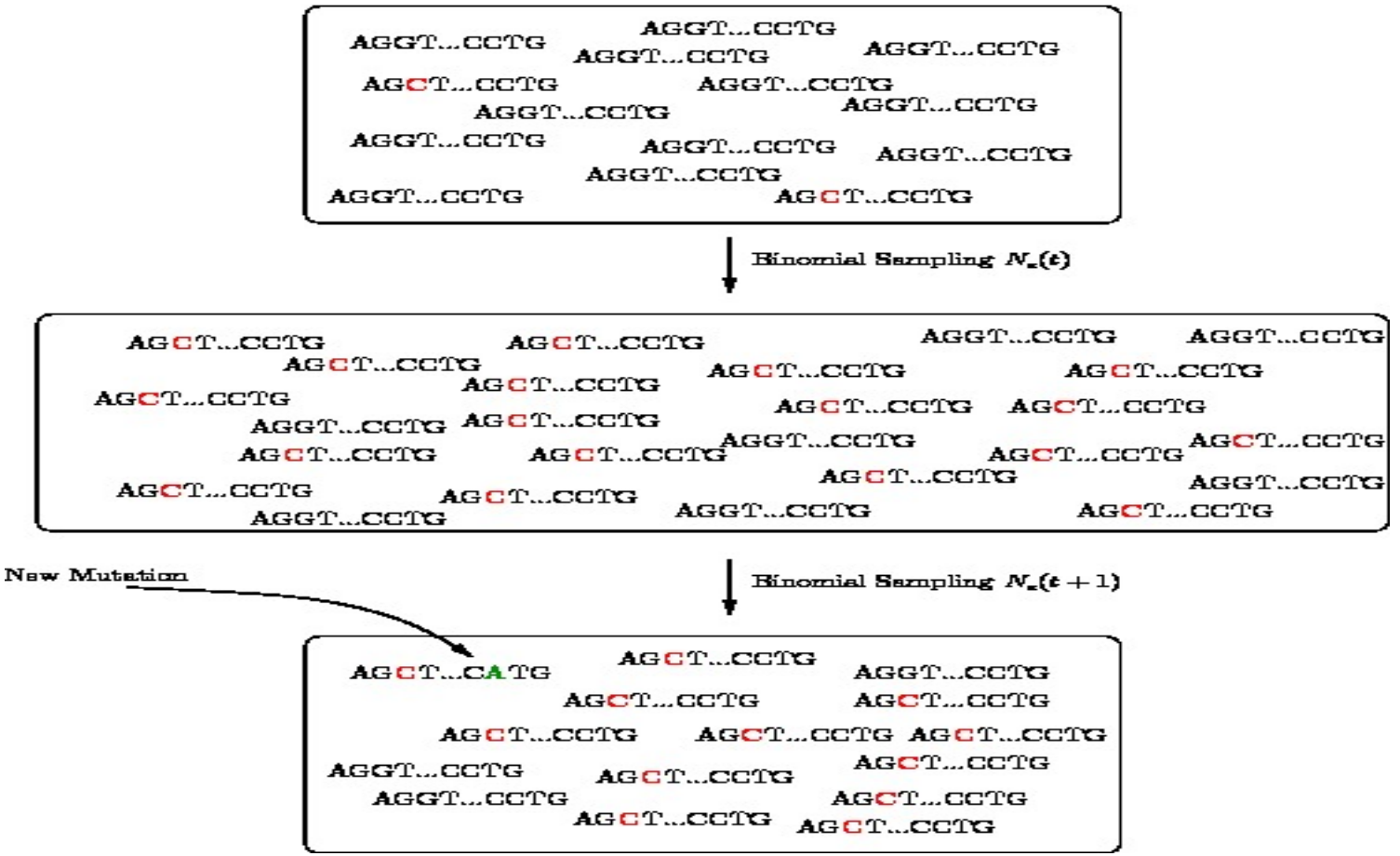


Pairwise difference

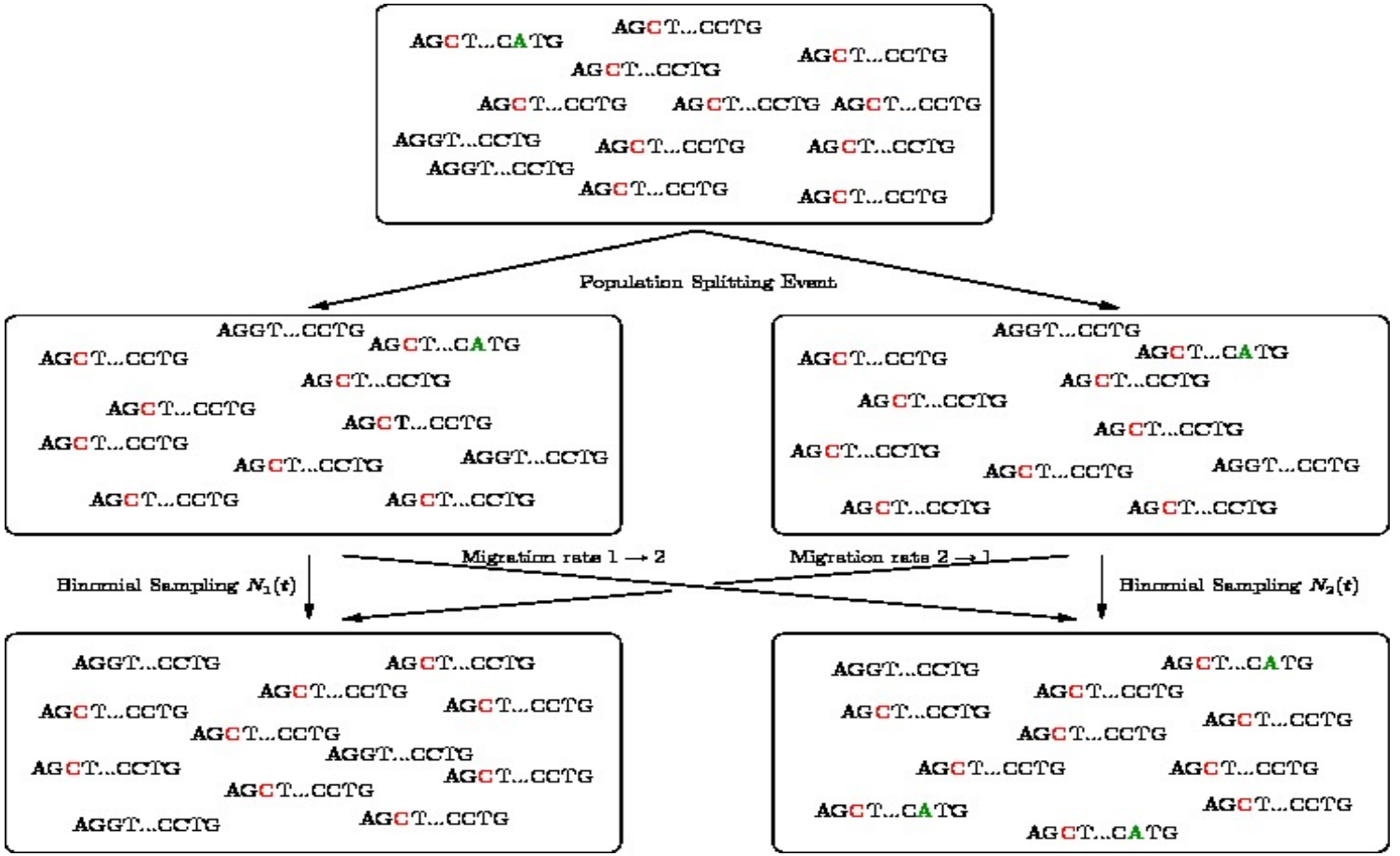
Boissinot et al. MBE (2001)

2.0. Diffusion Theory and Demography

The theory that describes the evolution of allele frequencies of diallelic markers in a population goes back to Fisher and Wright.



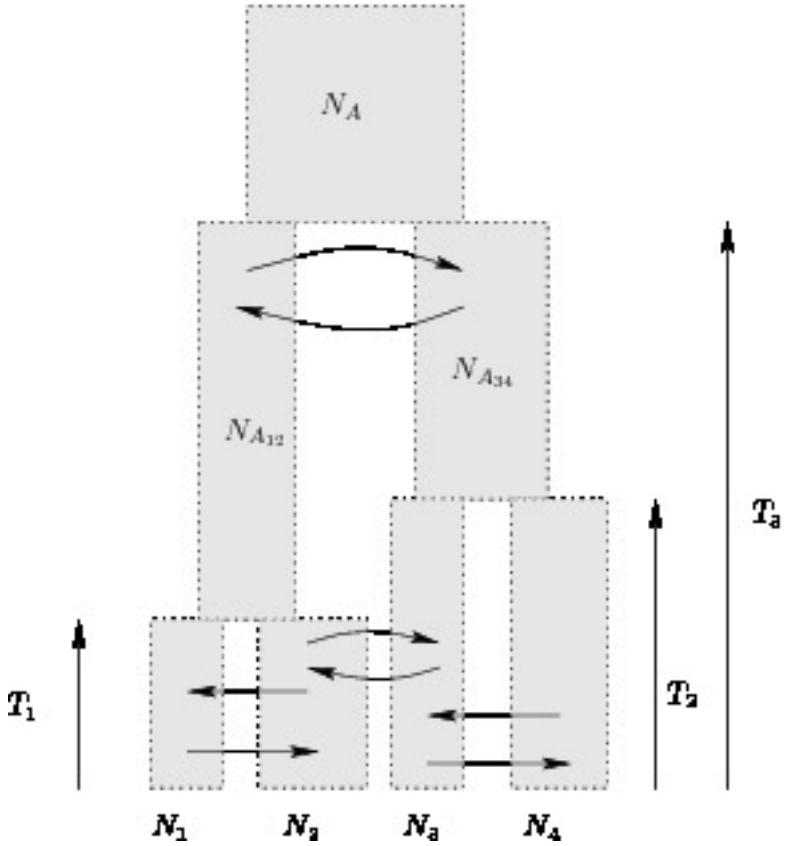
2.0. Diffusion Theory and Demography



2.0. Diffusion Theory and Demography

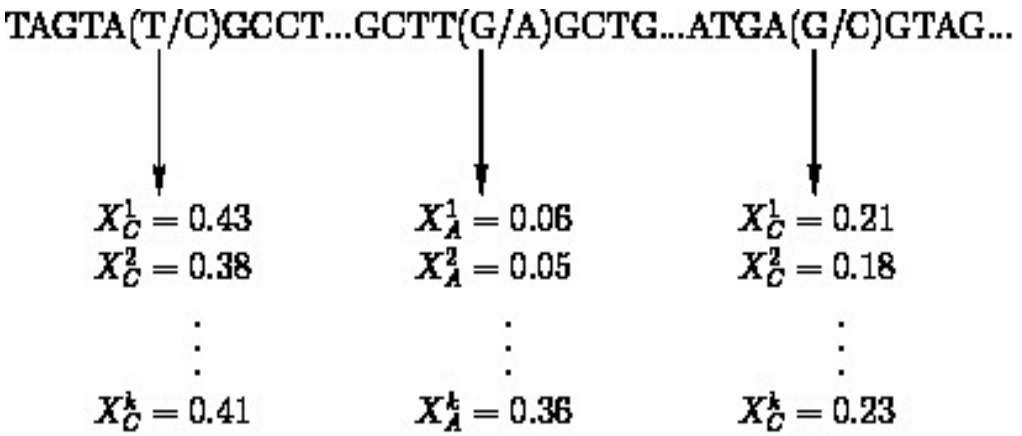
As a first approximation we consider models of random drift, migration between populations, influx of mutations and population splitting events.

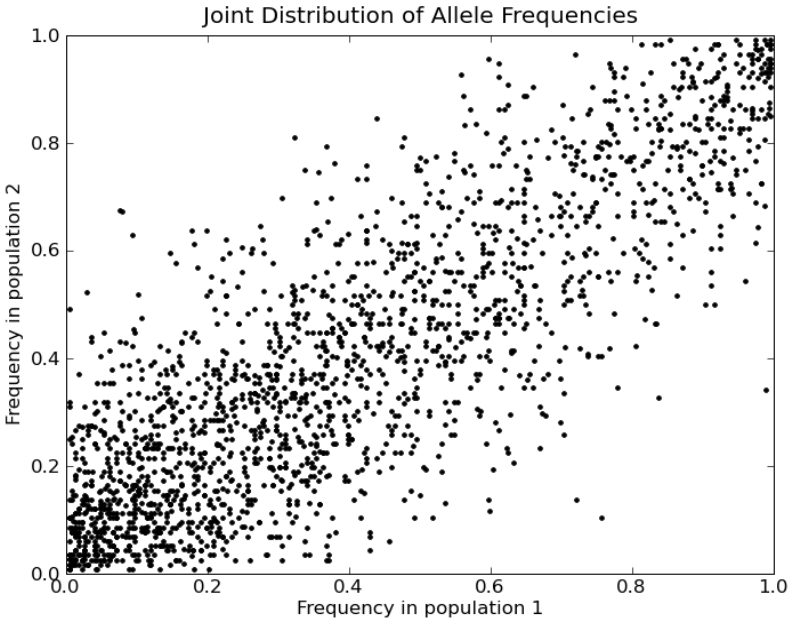
We assume free recombination between loci (no linkage).



2.1. Allele Frequency Spectrum

Given a particular population tree topology T and demographic parameters Θ, our model predicts particular Allele Frequency Spectra (AFS) that can be compared with the data.





2.1. Allele Frequency Spectrum

The AFS is the distribution of joint allele frequencies at the time when the samples were collected. This can be seen as a set of points in the k-cube, distributed according to certain probability density φ(x).

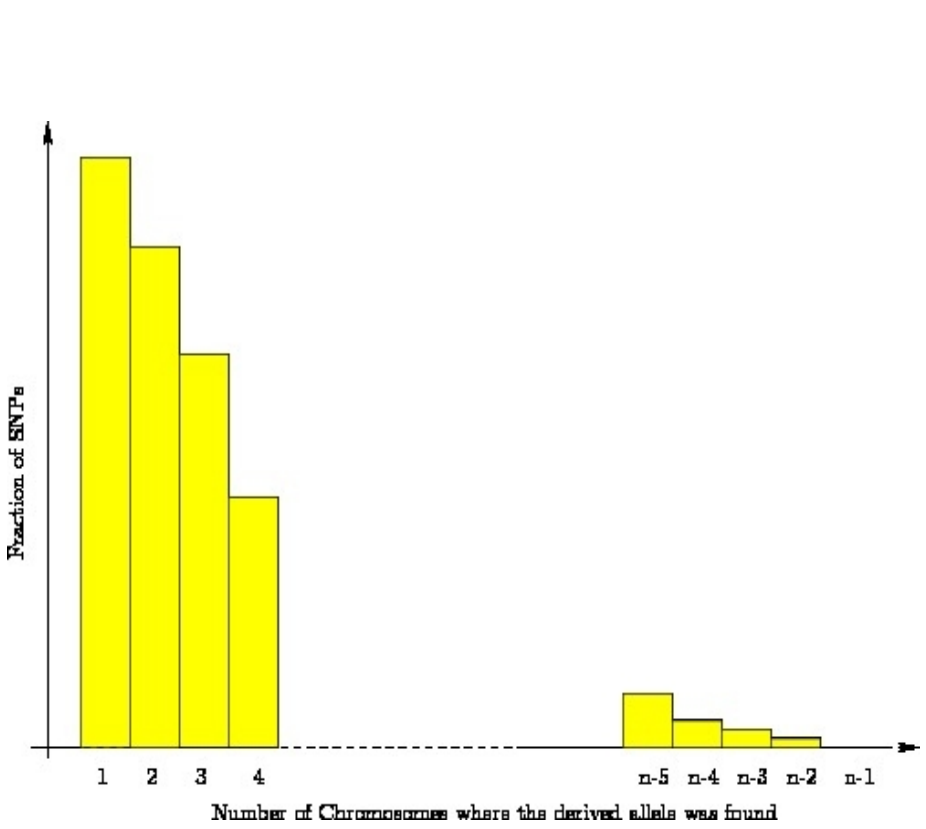
A *finite observation* of the joint AFS is a k-dimensional matrix with the allele counts in a finite sample.

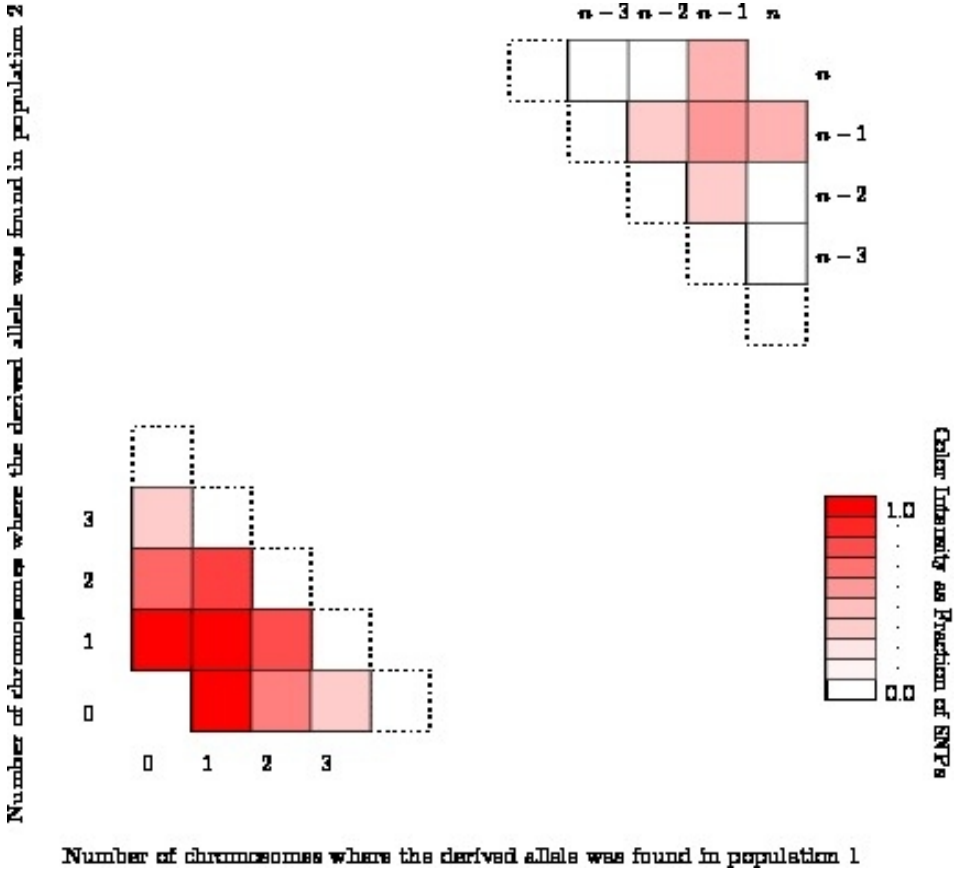
$$f_{i_1,i_2,\ldots i_k}(\Theta,T) = \frac{1}{Z}\int_{[0,1]^k} \phi(x|\Theta,T) \prod_{a=1}^{k} \frac{n_a!}{(n_a-i_a)!i_a!} x_a^{i_a}(1-x_a)^{n_a-i_a} dx_a.$$

$$Z = \sum_{0<\sum_a i_a<\sum_a n_a} \int_{[0,1]^k} \phi(x|\Theta,T) \prod_{a=1}^{k} \frac{n_a!}{(n_a-i_a)!i_a!} x_a^{i_a}(1-x_a)^{n_a-i_a} dx_a.$$

2.1. Allele Frequency Spectrum

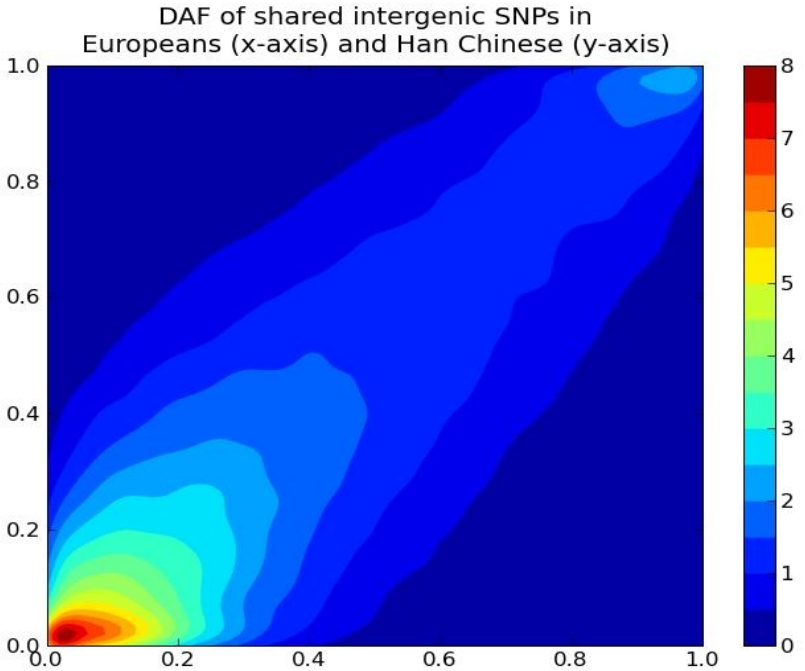
Examples of *finite observations* of AFS in one and two populations.

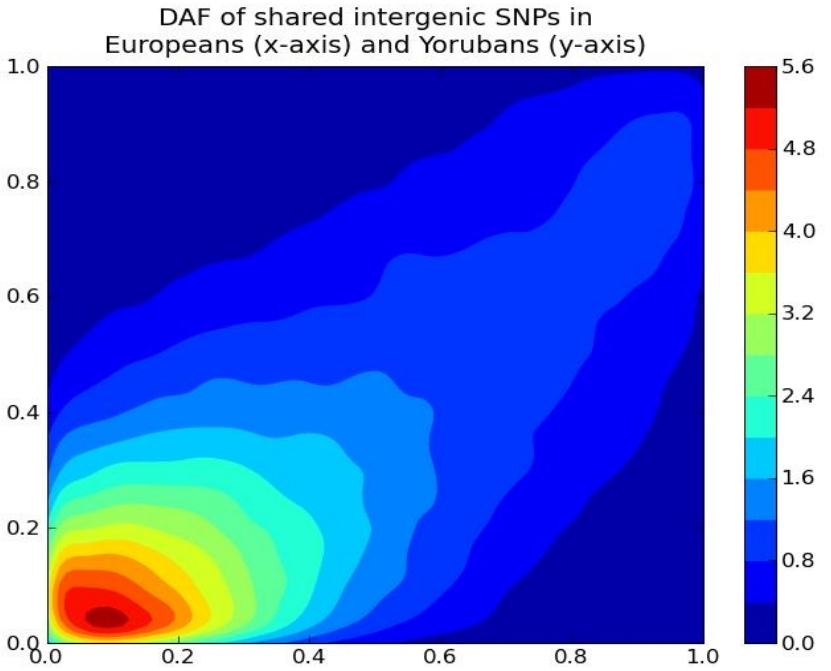




2.1. Allele Frequency Spectrum

A real example from the HapMap SNP data-set. (There are an average of 200 chromosomes per population.) Two 2d histograms with the density of derived alleles.



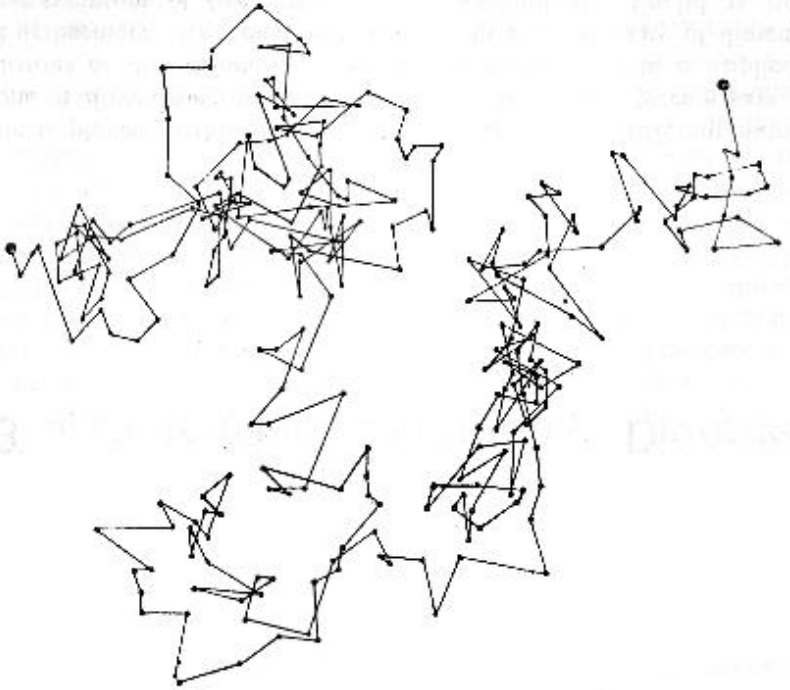


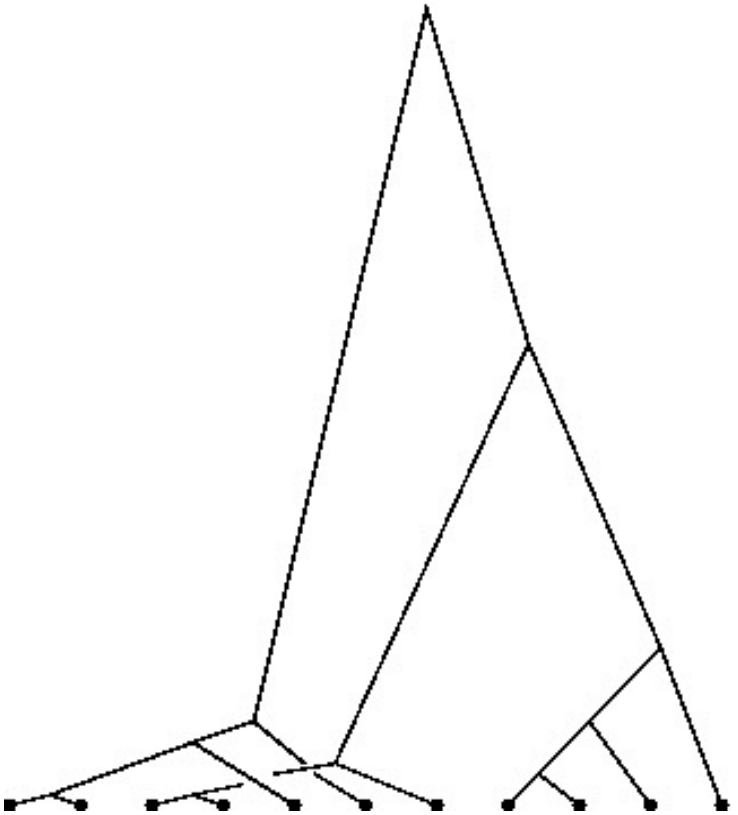
2.2. Inference of Demography from AFS

- Theory of predicting the AFS under irreversible mutation was developed by Fisher (1930), Wright (1938) and Kimura (1964). Kimura introduced the infinite sites model (1969).
- Wakeley-Hey (1997) computed the joint AFS using coalescent theory.
- Hudson (2002) computed the joint AFS using Monte-Carlo simulations of the diffusion process.
- Williamson et al. (2005), Evans et al. (2007), Gutenkunst et al. (2009), and S. L. et al. (2011) numerically solved the Partial Differential Equations (PDEs) associated with the infinite sites model.

2.2. Inference of Demography from AFS

Most popular approaches, such as coalescent-based simulations (trees within trees) and Monte-Carlo simulations, are so computationally intensive that complete investigations of the statistical properties of such models are limited to very simple cases (k=2).





2.2. Inference of Demography from AFS

Diffusion theory-based modeling allows thorough statistical studies of large data-sets and of families of demographic scenarios that involve more than two populations with migration.

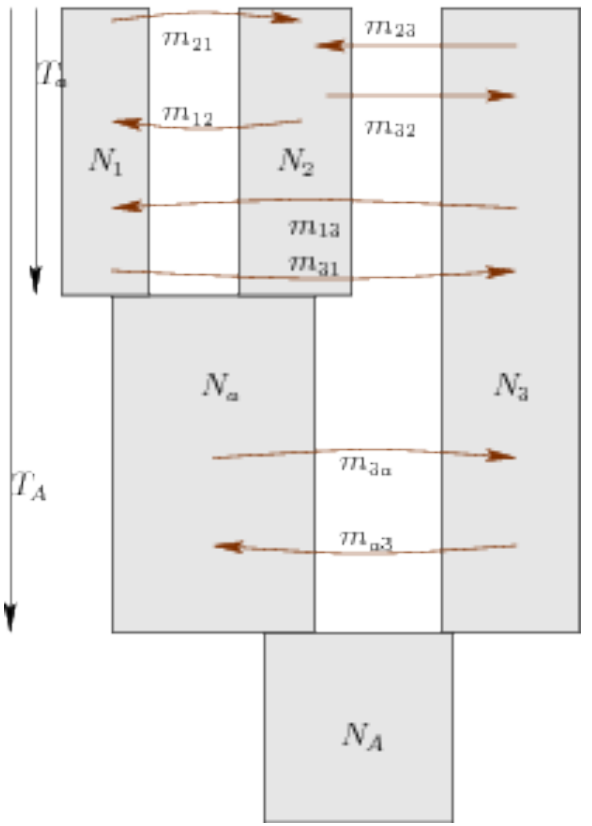
Another important advantage of the diffusion approach with respect to coalescent-based modeling is the ease with which selection can be incorporated.

3.0. Diffusion Approach to Demography

Solving the associated PDEs that model the evolutionary process requires the use of numerical approximations.

Special approximations are needed to deal with boundary conditions, influx of mutations, and population splitting events.

The two preferred methods to numerically solve the PDEs are *finite differences schemes* and *spectral methods*.

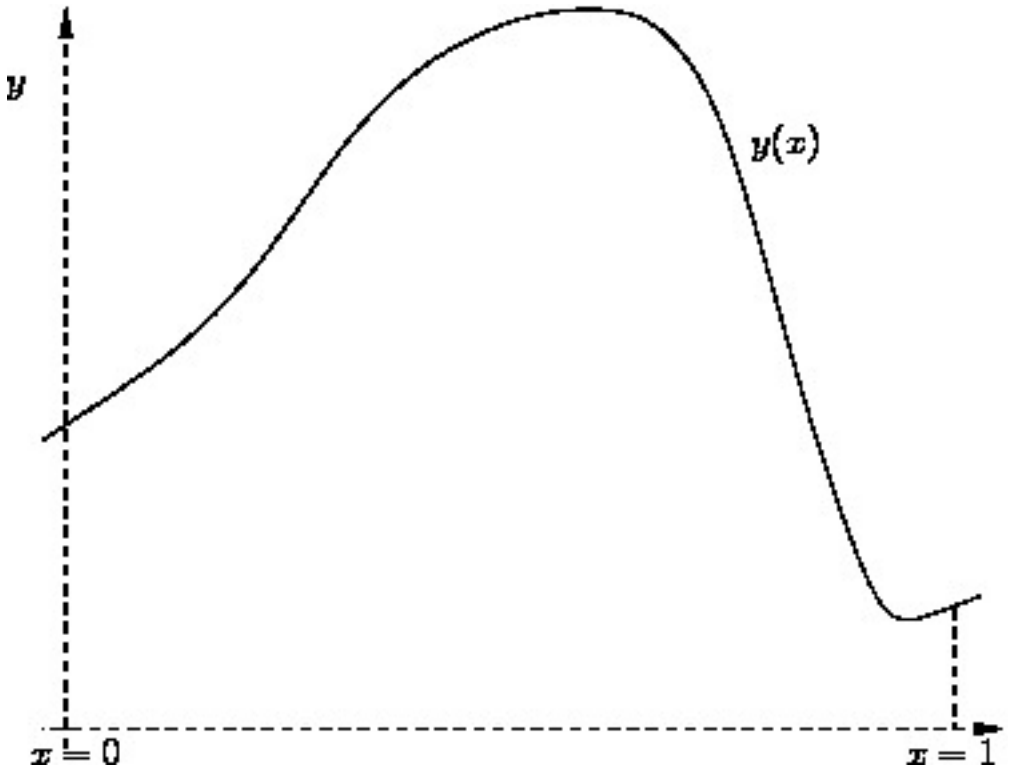


$$\frac{\partial}{\partial t}\phi(x,t) = \sum_{a,b}\frac{1}{2}\frac{\partial^2}{\partial x_a \partial x_b}\left(\delta^{ab}\frac{x_a(1-x_a)}{2N_{e,a}}\phi(x,t)\right) - \frac{\partial}{\partial x_a}\left(m_{ab}(x_b - x_a)\phi(x,t)\right) + \rho(x,t)$$

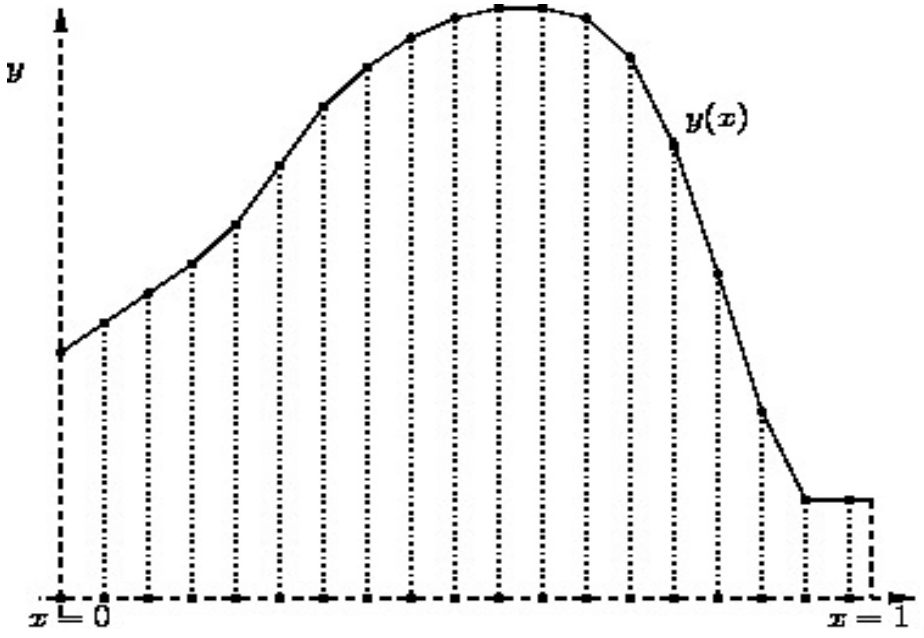
3.0. Diffusion Approach to Demography

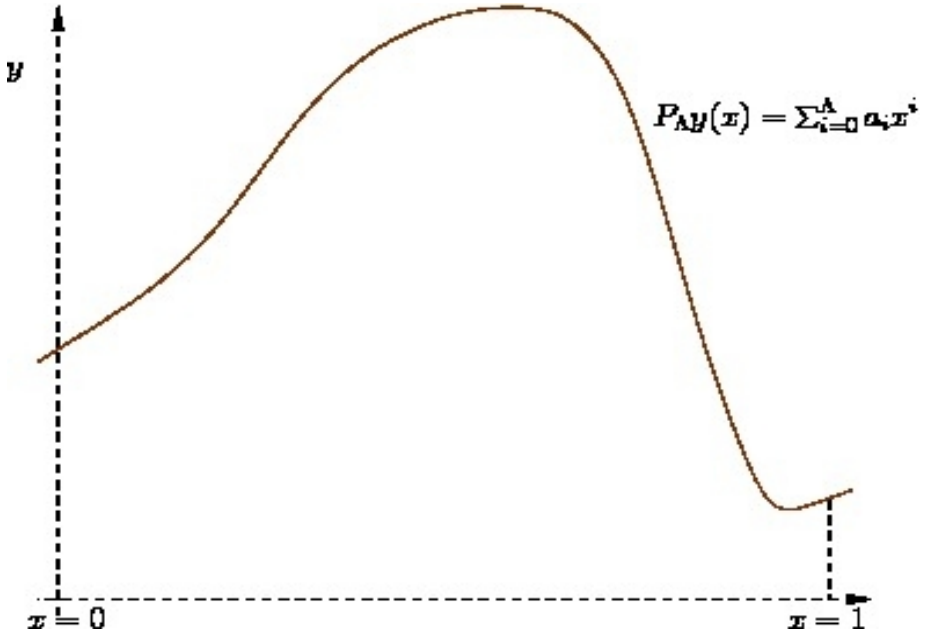
Finite differences schemes are very robust and simple to implement, permitting the analysis of complex demographic scenarios with up to three simultaneous populations with migration.

Spectral methods are more difficult to implement, although they are usually the preferred methods when *the dimension of the domain is high* and the solutions of the diffusion PDEs are *smooth*.



3.1. Numerical Solutions to the PDEs

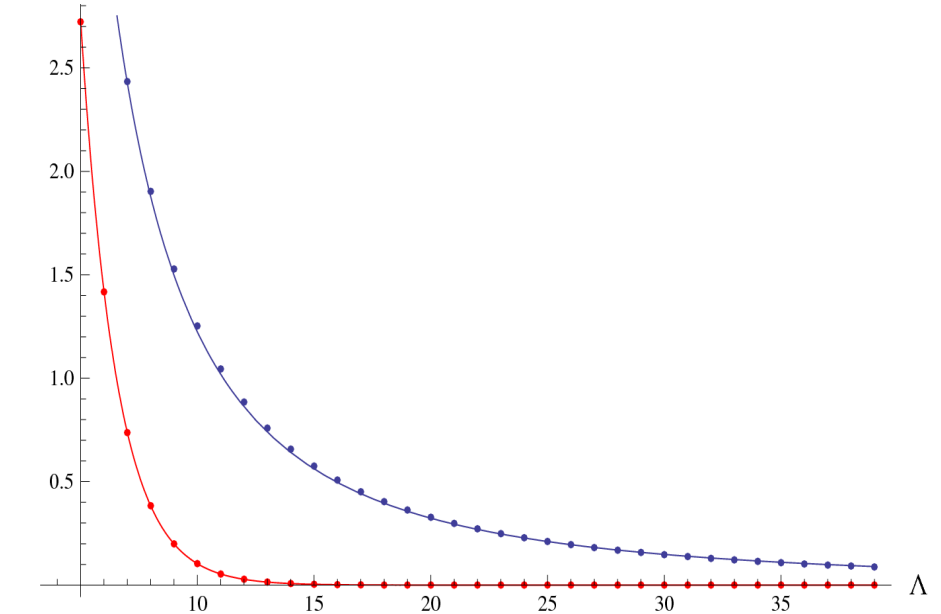




$$\text{Error} = \int_0^1 |y(x) - y_{app}(x)| dx$$

$$101.85\Lambda^{-1.92}.$$

$$72.16e^{-0.65\Lambda}.$$



3.1. Numerical Solutions to the PDEs

We write the solution to the PDE as truncated polynomial expansion:

$$\phi(x,t) = \sum_{i_1=0}^{\Lambda_1-1} \sum_{i_2=0}^{\Lambda_2-1} \cdots \sum_{i_P=0}^{\Lambda_P-1} \alpha_{i_1,i_2,\ldots i_P}(t) R_{i_1}(x_1) R_{i_2}(x_2) \cdots R_{i_P}(x_P).$$

This gives a finite dimensional approximation to the spaces of densities that scales as Λ^P. In a finite differences scheme, the finite dimensional approximation to the spaces of densities scales as M^P, with $M >> \Lambda$ working at the same level of accuracy.

We project the diffusion PDE in the Fourier space, and solve the associated ODE.

$$\frac{\partial \alpha_I(t)}{\partial t} = \sum_J \omega_{IJ}(t)\alpha_J(t) + \beta_I.$$

3.1. Numerical Solutions to the PDEs

Singularities that appear when modeling population splitting events (jumping from dimension k to dimension k+1), can be substituted by smoothed approximations.

The influx of mutations, which is defined by a Dirac delta in Kimura's classical exact solution

$$\mathcal{G}(x;p,t) = \sum_{i=0}^{\infty} \frac{2(i+1)+1}{(i+1)(i+2)} (1-(1-2p)^2) T_i(1-2p) T_i(1-2x) \exp(-(i+1)(i+2)t/4N_e)$$

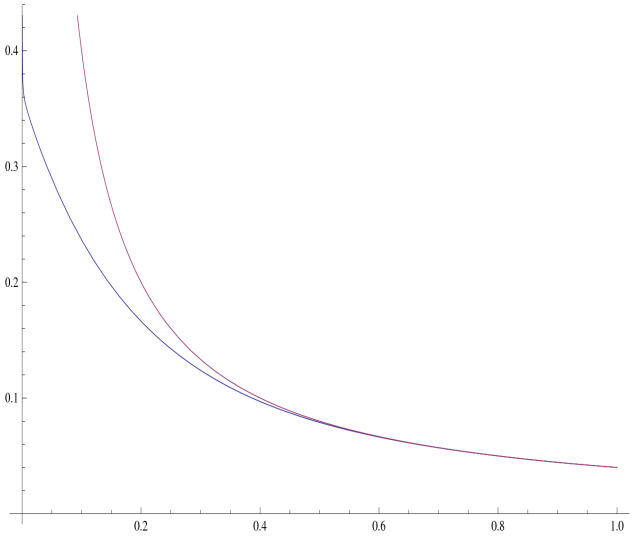
$$\phi(x,t) = \int_0^1 \mathcal{G}(x;y,t)\phi(y,t=0)dy + 2N_e u \int_0^t \mathcal{G}(x;1/2N_e,\tau)d\tau$$

can be substituted by a more general smooth *effective mutation density*. Both solutions, converge to the same AFS density (S.L. et. al. 2011).

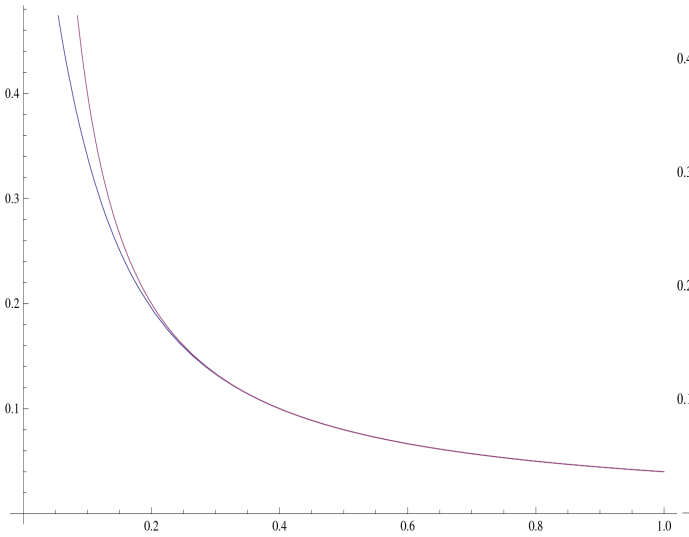
3.1. Numerical Solutions to the PDEs

Comparison of the equilibrium densities of Derived Alleles in one population, associated to an *effective mutation density* and to the standard model:

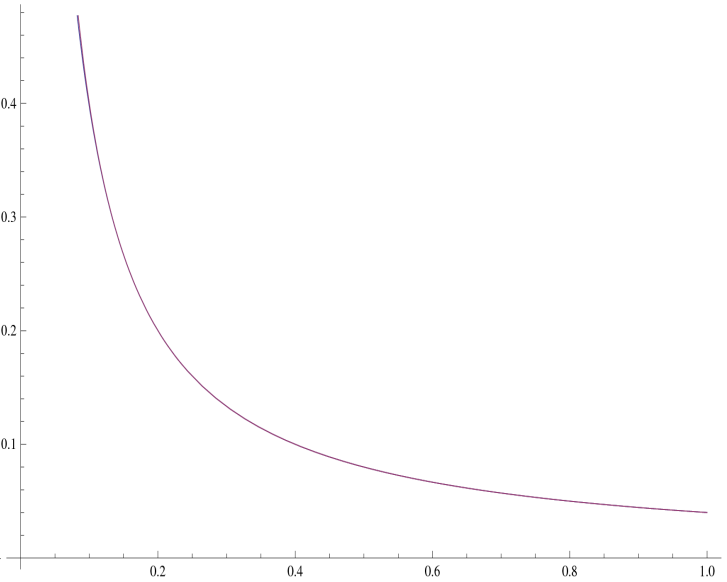
Λ=6



Λ=12



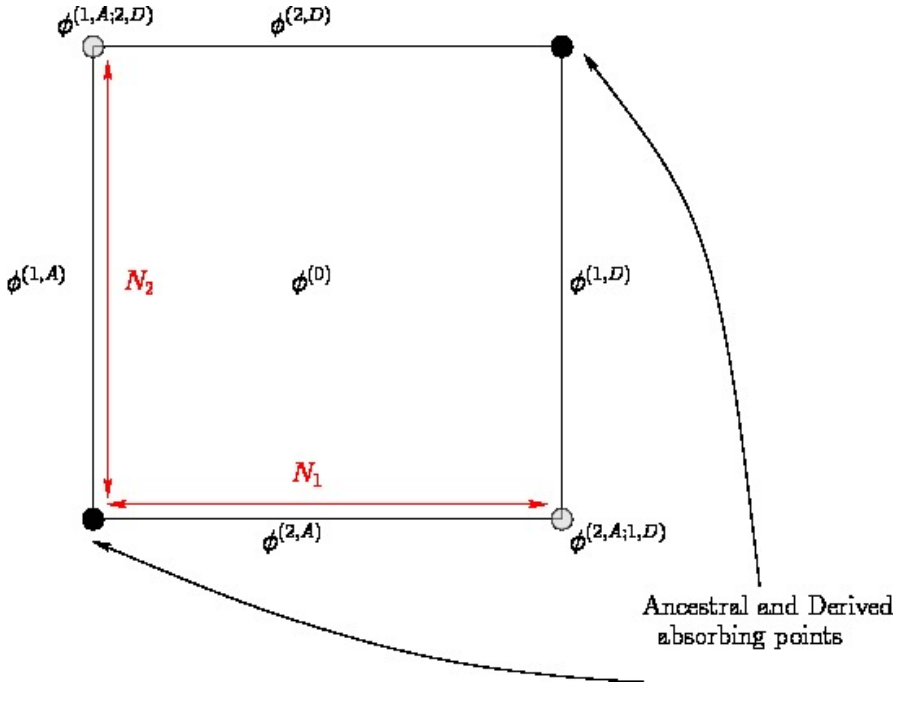
Λ=18

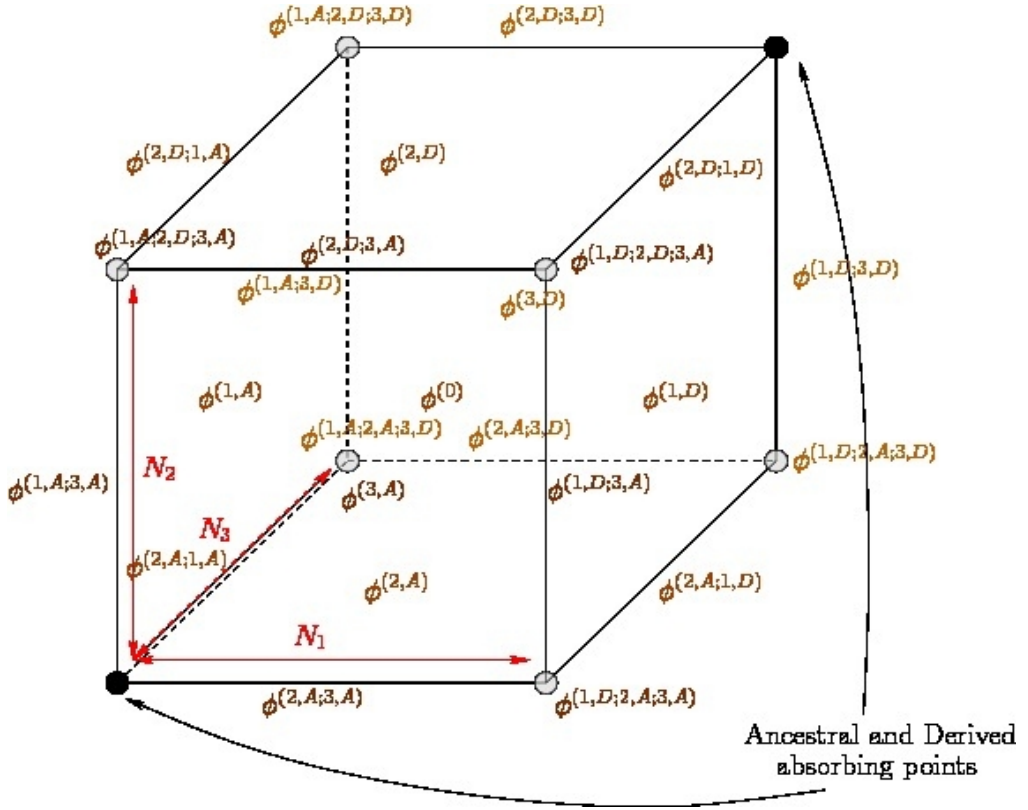


3.1. Numerical Solutions to the PDEs

The contribution to the density from each boundary component can be approximated by a smooth function.

Each term interacts with higher and/or lower boundary components via fixation of alleles in certain populations, migration events, and influx of mutations (S.L. et al. 2011)





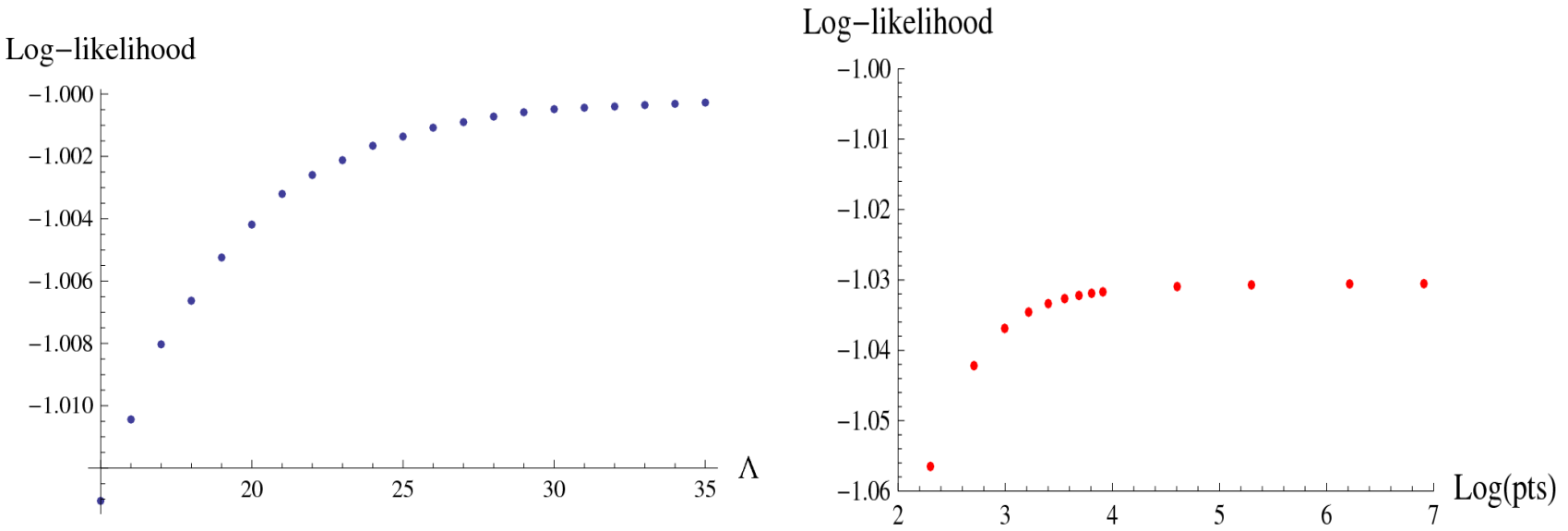
3.1. Numerical Solutions to the PDEs

We have shown how to use truncated polynomial expansions to solve diffusion processes on trees that model influx of mutations, random drift, migration and multiple population splitting events.

Despite the fact that polynomial expansions fail to approximate non-smooth functions (e.g. Dirac distributions), we have shown how every event that happens in natural evolutionary histories of populations can be approximated by smooth functions.

Given a tree topology T and model parameters θ, this approach yields accurate and fast evaluations of AFS, and comparisons with observed AFS.

4.0. Comparison with other methods

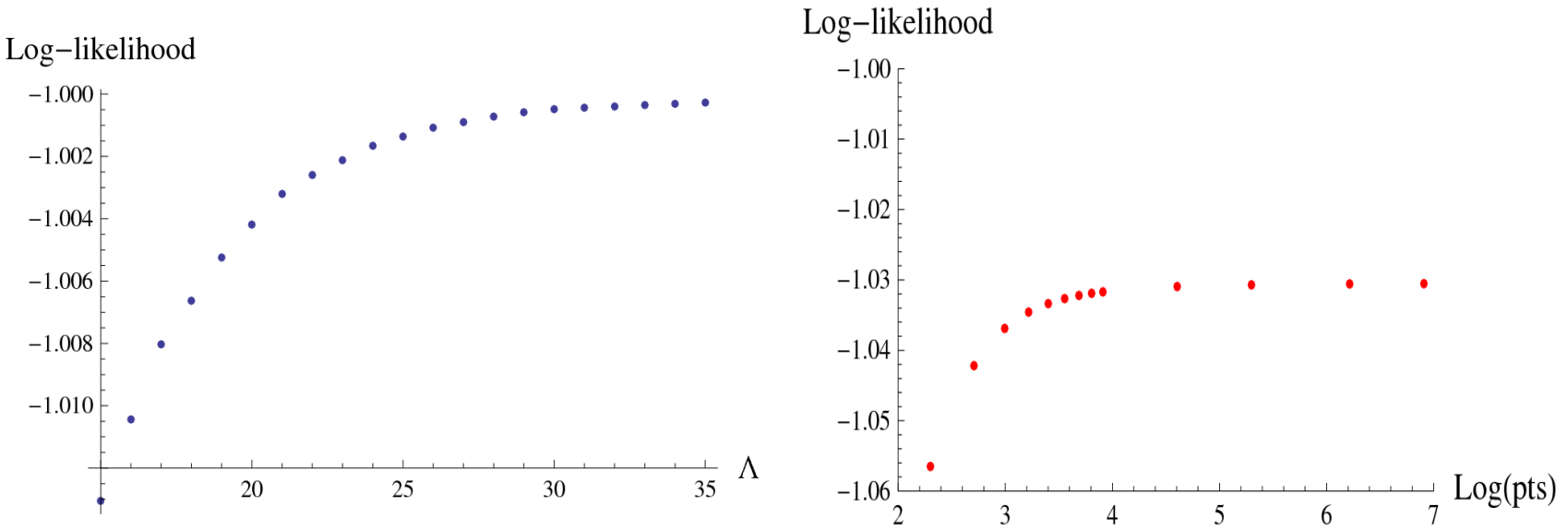


We generated simulated data given different demographic scenarios with 2 and 3 populations.

We compared our implementation (left) with DADI (right).

DADI uses a Crank-Nicholson finite difference method, with an unknown mutational model, and zero flux at the boundary.

4.0. Comparison with other methods



Our method is slower than DADI if run on a single CPU (we use less memory and more CPU).

DADI gives fast approximations to the exact solutions of the PDE but does not converge to it (p-value~0.001).

Our method does converge (p-value~0.3). This might be due to our mutational model and choice of boundary conditions.

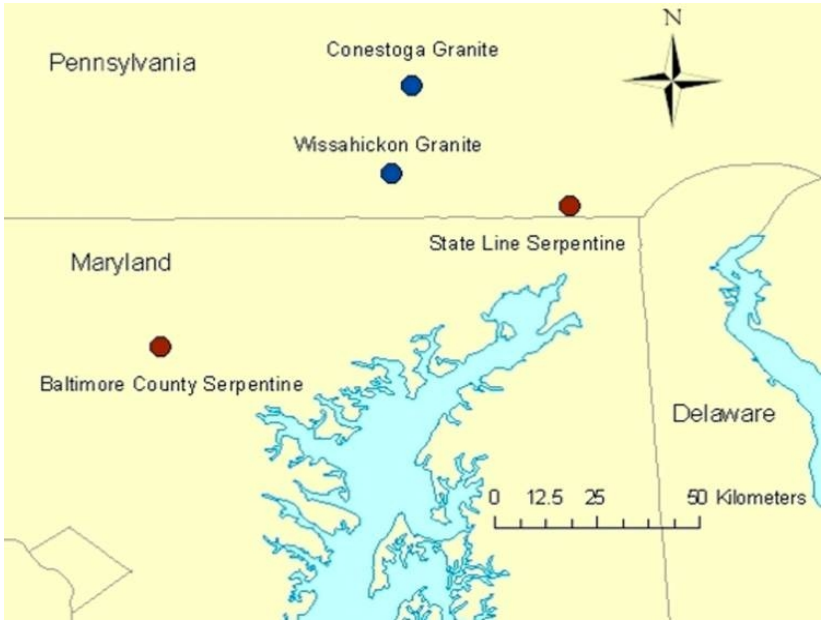
4.1. Example: 4 Populations of A. lyrata

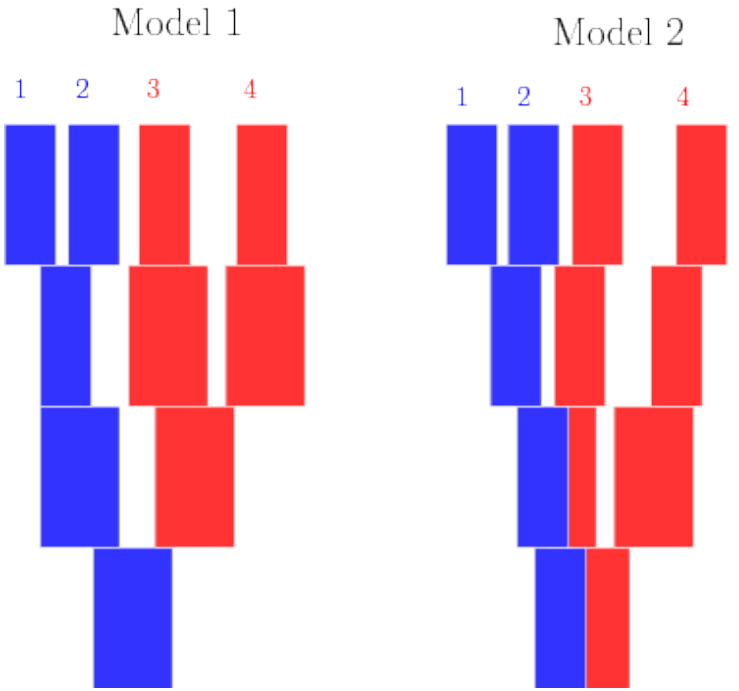
Arabidopsis lyrata is an outcrossing diploid with a small genome and a non-trivial demographic history linked to local adaptations of its populations. (T. L. Turner et. al. Nature Genetics 2010).

AFS density shaped by demography and adaptation.

A. lyrata's proximity to the genetic model organism A. thaliana allows us to identify neutral polymorphisms.

4.1. Example: 4 Populations of A. lyrata





1=Conestoga granite; 2=Wissahickon granite; 3=State Line serpentine; 4=Baltimore county serpentine.

In red we denote the populations found in serpentine soils; in blue we denote those found in granitic soils.

25 diploid individuals per population were genotyped.

4.1. Example: 4 Populations of A. lyrata

After rooting with A. thaliana to identify ancestral states, 8,433,201 SNPs were detected.

5,465,168 SNPs were annotated as non-coding in A. lyrata.

Of those, ~300,000 SNPs were polymorphic in the 4 populations simultaneously.

High-throughput sequencing of DNA pooled from the 25 individuals, with 39-fold coverage, was the source of data (T. L. Turner, Nature Genetics 2010).

For each diallelic SNP, and for each population we have two observations: Total number of counts R, and number of counts of the derived allele r.

4.1. Example: 4 Populations of A. lyrata

Given a SNP, with total number of counts R, number of counts of the derived allele r, and hidden number of derived alleles i, the observed AFS will be given as

$$p(r|i,\, R,\, n) = \frac{R!}{r!(R-r)!}\left(\frac{i}{n}\right)^{r}\left(1-\frac{i}{n}\right)^{R-r}$$

$$p(i|r,\, R,\, n) = \frac{(i/n)^{r}(1-i/n)^{R-r}}{\sum_{j=0}^{j=n}(j/n)^{r}(1-j/n)^{R-r}}$$

$$m_{i_1,i_2,i_3,i_4} = \sum_{q\in Q} p(i_1|r_1^q,\, R_1^q,\, n)p(i_2|r_2^q,\, R_2^q,\, n)p(i_3|r_3^q,\, R_3^q,\, n)p(i_4|r_4^q,\, R_4^q,\, n)$$

$$\log-\text{likelihood} = \sum_{i_1,\ldots i_k} m_{i_1,i_2,\ldots,i_k}\log\left(f_{i_1,i_2,\ldots,i_k}(\Theta,T)\right)$$

4.1. Example: 4 Populations of A. lyrata

The diffusion model with migration and 3 splitting events has 49 free parameters. We denote the space of model parameters as Ω.

We use a C++ implementation of the regularized setup to the problem here presented to solve the diffusion equations. The code will be freely available soon.

We compute the likelihood given the data by considering expectations for smaller sample sizes (10x10x10x10), on the folded AFS.

4.1. Example: 4 Populations of A. lyrata

Solving the diffusion equations for each point $\theta \in \Omega$ takes between 125 secs and 200 secs depending on θ, using a single CPU and very little memory.

Maximum Likelihood Estimate is found by using the BFGS Quasi-Newton and down hill simplex methods.

Importance Sampling (using an auxiliary density given as a mixture of Gaussians on the parameter space) is used to describe the posterior probability distribution on the parameter space of the model given the data.

4.1. Example: 4 Populations of A. lyrata

u=3e-8 b.s. per gen.

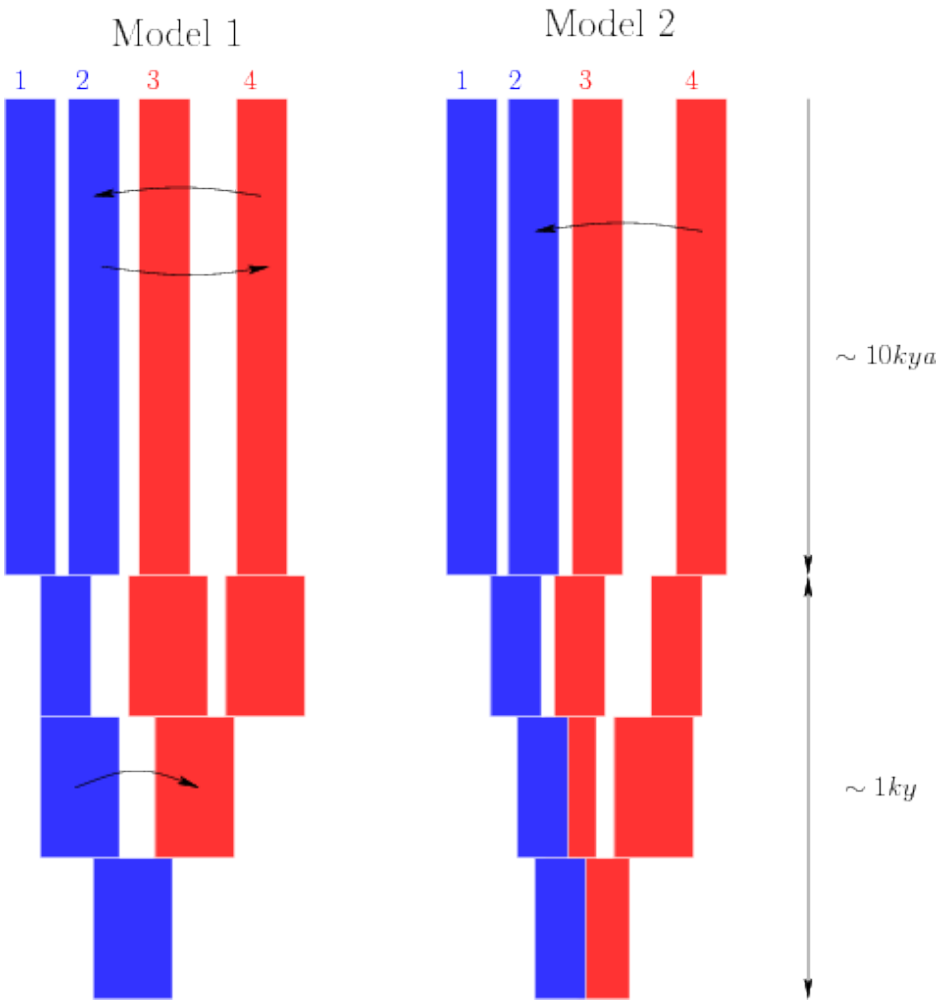
N_1=12,500-14,000

N_2=11,700-13,300

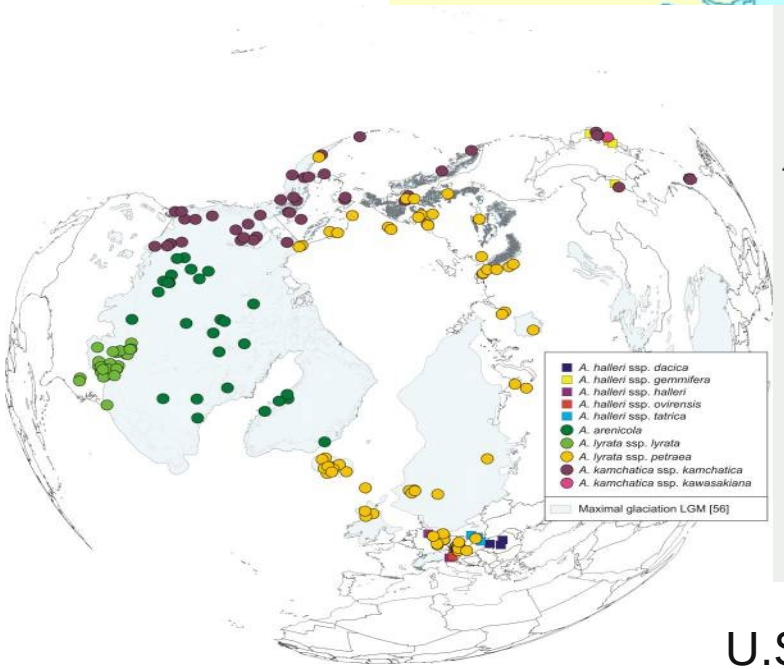
N_3=12,700-14,400

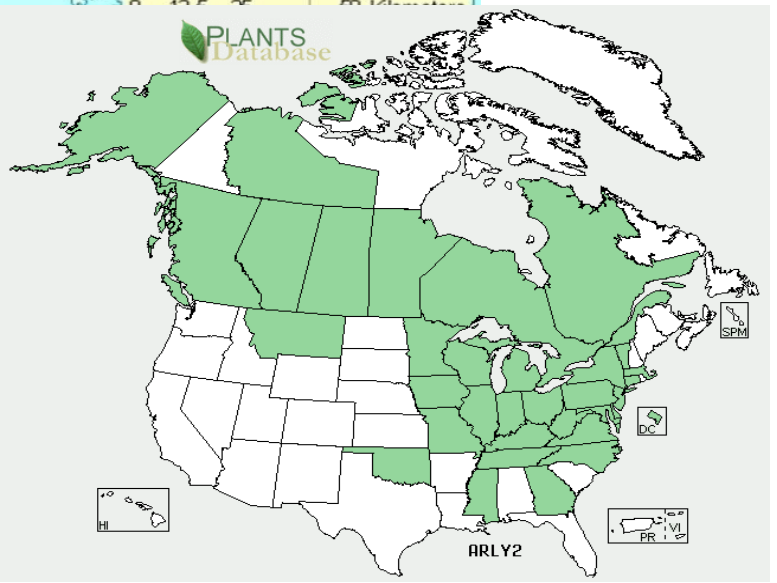
N_4=7,800-8,900

Gene-flow: 0.2 haploid genomes per gen. Migration rates with 2Nm<0.1 are ignored in the figure.



4.1. Example: 4 Populations of A. lyrata

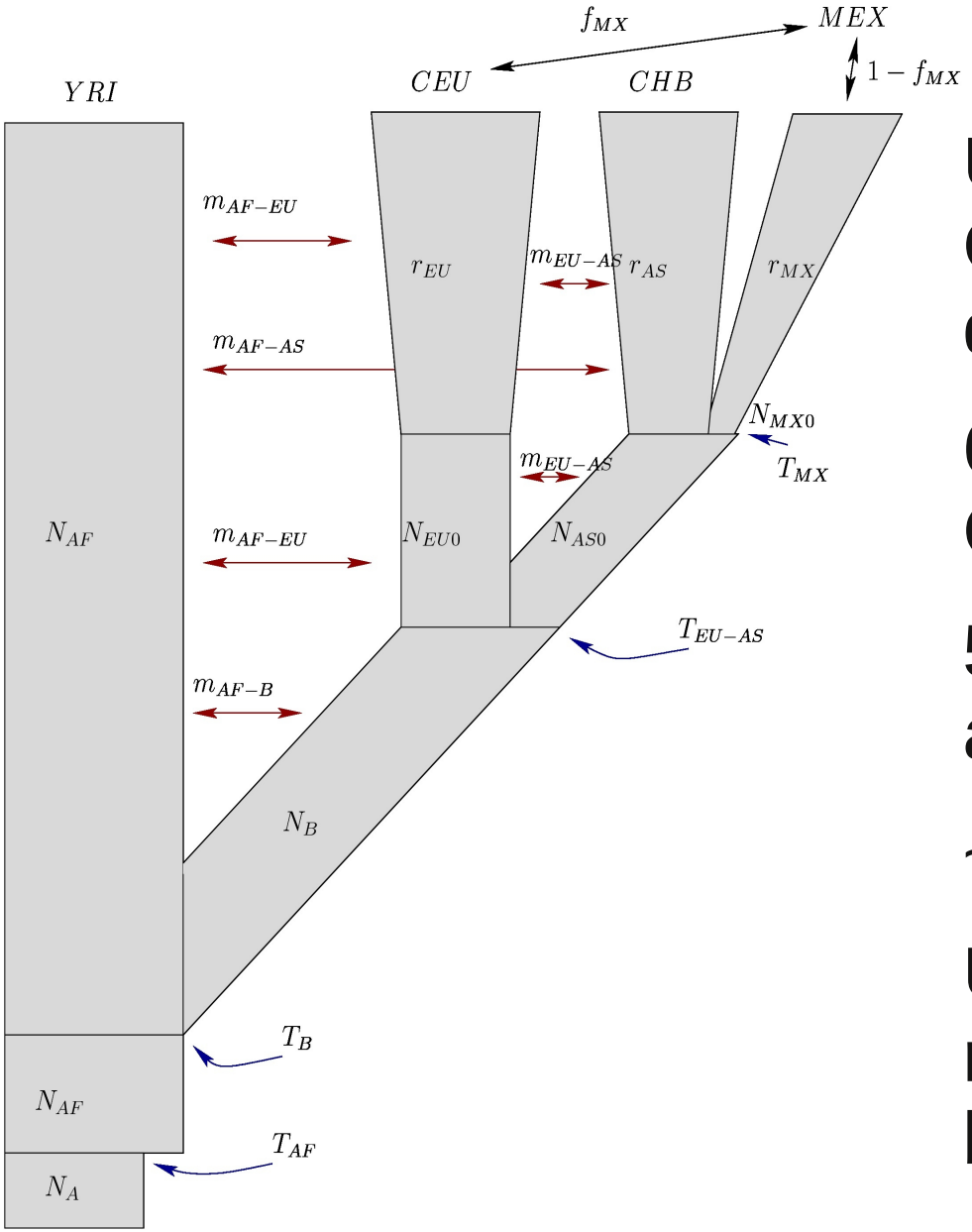




U.S. Department of Agriculture webpage.

R. Schmickl et al. BMC Evol Biol. 2010

4.2. Example: Human expansion out of Africa and settlement of America



Using Environmental Genome Project sequence data.

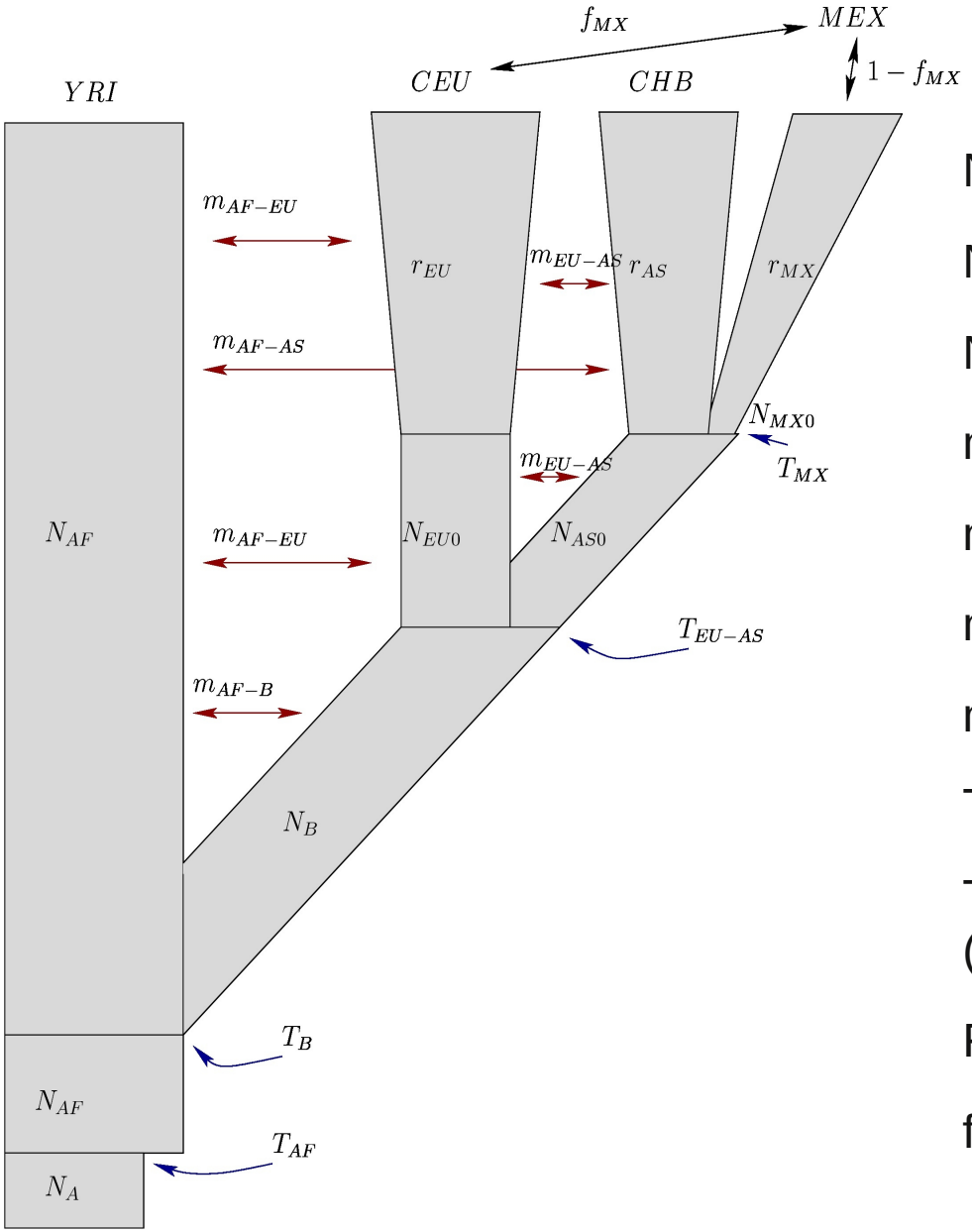
68 individuals (YRI, CEU, CHB, MEX).

5.0 Mb of DNA resequenced at very high coverage.

~26,000 SNPs.

Used an adaptive AFS to reduce computational burden.

4.2. Example: Human expansion out of Africa and settlement of America



N_{AF}=16500 (1500) N_B=1200 (300)

N_{EU0}=1800.0 (400) r_{EU}=0.62 (0.10)

N_{AS0}=5100 (1000) r_{AS0}=0.0048 (0.002)

m_{AF-B}=3.5e-05 (1.5e-05)

m_{AF-EU}=1.6e-05 (1.0e-05)

m_{AF-AS}=6.7e-05 (2.0e-05)

m_{EU-AS}=4.0e-05 (1.5e-05)

T_{AF}=270kya (80) T_B=20kya (5)

T_{EU-AS}=17.2kya (4.5) N_{MX0}=2100 (1100)

R_{MX}=0.56 (0.21) T_{MX}=17.1kya (4.5)

f_{MX}=0.32 (0.10)

Conclusions

We have introduced a particular regularization of the influx of mutations, population splitting events and the boundary conditions, that allow us to solve numerically the PDEs associated with the multiple population infinite sites model using very efficient numerical methods.

With these methods we can work easily with four populations and find MLE for models with a large number of parameters.

Thank you!

- Jody Hey.
- Kevin Chen.